Contribution de la métrique à la stylométrie

V. Beaudouin, F. Yvon

Beaudouin, Valérie & Yvon, François (2004). *Contribution de la métrique à la stylométrie*. JADT 2004 - Le poids des mots, Louvain, UCL - Presses Universitaires de Louvain, p. 107-118.

Contribution de la métrique à la stylométrie

Valérie Beaudouin[1] et François Yvon[2]

[1]FT R&D, 38-40 rue du Général Leclerc, Issy les Moulineaux

[2]GET/ENST et CNRS/LTCI– 4- rue Barrault– 75013 Paris – France

Abstract

In this paper, we present the results of our first attempts to use metrical informations for authorship attribution. 58 plays in verse by Corneille, Racine and Molière were systematically analyzed regarding syllabic, morphosyntactic and stress structure, using the Metrometer, a Natural Language Processing tool aimed at the analysis of French classical verse. Our main finding are that (i) Corneille and Racine are very consistent in their metrical use of words, Molière being quite less so; (ii) using statistical language modeling tools, it is in fact possible to build syllable-based model which can effectively discriminate genre and authors.

Résumé

Nous montrons comment des informations métriques du vers peuvent être utilisées dans le cadre d'études stylométriques. L'ensemble des pièces en vers de Corneille, Racine et Molière, soit 58 pièces, a été analysé par le Métromètre, un outil d'analyse du vers classique qui produit pour chaque position métrique un ensemble de descriptifs linguistiques (syllabe, catégorie morpho-syntaxique, accent…). Alors que Corneille et Racine sont cohérents à travers leur oeuvre dans leur manière de procéder aux décomptes métriques dans le vers, Molière est quant à lui plus fluctuant, au moins sur trois mots. Toutefois, lorsqu'on les considère de manière indépendante, les différents descripteurs des positions métriques utilisés ici ne permettent pas de différencier de manière sûre des auteurs pour un genre donné. En revanche, l'utilisation de modèles statistiques du vers, fondés sur la séquence syllabique, permet de construire des outils de discrimination qui conduisent à identifier avec une précision raisonnable genres et auteurs.

Mots-clés : Stylométrie, Modèles de langage.

Les questions d'attribution des textes ont régulièrement suscité des débats passionnels, comme les travaux anglo-saxons sur les textes de Shakespeare ou les récentes études sur Corneille et Racine (Labbé et Labbé, 2001). Les polémiques s'intensifient avec la notoriété des auteurs mis en jeu : changer l'attribution de pièces classiques revient à mettre en péril des savoirs qui s'appuient souvent sur les biographies des auteurs pour comprendre les œuvres. L'intérêt de la sphère médiatique pour ces remises en cause tend à aggraver les polémiques. Les travaux de stylométrie qui traitent des questions d'attribution dérangent car ils mobilisent des techniques encore peu répandues dans le milieu littéraire et souvent font abstraction de toute la connaissance accumulée dans le domaine, passent parfois rapidement sur des notions aussi centrales que celle de genre. Les débats sont également intenses entre les chercheurs du domaine, comme en témoigne la synthèse qu'Holmes (1998) propose des travaux en stylométrie.

Différents types d'indicateurs ont été utilisés dans les travaux d'attribution (mots les plus fréquents, mots outils, rimes….) (Holmes, 1998). Nous proposons de tester les éléments métriques (syllabes, catégories morpho-syntaxiques et accents) comme candidats potentiels à une différenciation des auteurs et des genres. Y a-t-il une manière propre à chaque auteur de faire des vers ? En quoi les aspects métriques peuvent-ils qualifier l'écriture? Nous nous appuyons sur les pièces en vers de Corneille, Molière et Racine[i] en accordant une place particulière aux comédies, puisque les différences de genre l'emportent généralement sur les écarts entre auteurs. Les aspects métriques sont traités avec un outil, le métromètre, mis au point en 1993 pour l'analyse du vers classique (Beaudouin et Yvon, 1995).

Après une brève présentation du corpus utilisé et du métromètre, nous montrons que le traitement métrique de certains mots dans le vers peut permettre de distinguer les auteurs entre eux. Ces indices très ténus doivent être complétés par une approche plus globale du vers. Après avoir montré que des statistiques descriptives sur la plupart des critères métriques ne permettent pas de distinguer auteurs et genres, nous recourons aux modèles de langage appliqués aux séquences de syllabes métriques. Des modèles sont construits pour les auteurs et les genres, sur des échantillons de pièces, et on teste la capacité du système à attribuer les vers restant au bon modèle.

1. Corpus et métromètre

Le corpus est constitué par l'ensemble des pièces en vers de Corneille, Molière et Racine, soit 58 pièces.

	Comédies	Tragédies	Pièces diverses
Corneille	9	21	4
Racine	1	11	-
Molière	12	-	-

Tableau 1 : Répartition des pièces du corpus par auteur et genre

Les éditions électroniques utilisées correspondent à l'édition de Marty-Laveaux (1862) pour Corneille – qui intègre toutes les corrections faites en 1660 par Corneille sur ses premières pièces, à celle de Paul Mesnard (1885) pour Racine et à celle d'Eugène Despois (1873) pour Molière[ii].

Le métromètre est un outil de description systématique de la structure phonétique, morpho-syntaxique et accentuelle du vers. Il repose sur une analyse phonétique et métrique du vers. Cet outil résulte de l'adaptation au cas particulier du vers d'un phonétiseur du français développé par François Yvon (1995). Ce phonétiseur d'appuie lui-même sur un analyseur syntaxique, Sylex, développé par Patrick Constant (1991).

[i] Les œuvres de Corneille et Racine ont été les premières à être analysées en France avec les méthodes de statistique lexicale, développées par Charles Muller (Muller, 1967 ; Bernet, 1983). Ces travaux de lexicométrie ont permis entre autre de mettre en évidence les spécificités lexicales liées aux genres, aux auteurs et aux périodes.

[ii] Le corpus Molière nous a été fourni par Charles Bernet, que nous remercions. Cette édition électronique des pièces a été mise en place dans le cadre d'un projet INALF sur le théâtre du XVIIéme : les pièces y sont balisées dans un langage apparenté au XML, ce qui permet une préparation accélérée pour le passage au métromètre.

Le métromètre, après analyse syntaxique, transcrit le vers dans l'alphabet phonétique, le découpe selon les positions métriques, en respectant les règles de la versification (diérèse/synérèse, décompte du *e* muet et liaison) et finalement, attribue à chacune des positions un certain nombre de marquages ou étiquettes d'ordre exclusivement linguistique (syllabe et voyelle métriques, fin de mot, catégorie morpho-syntaxique, accent). Voici ce que donne l'analyse du vers suivant de Racine :

De cette nuit, Phénice, as-tu vu la splendeur ?

Racine, *Bérénice*, vers 302.

Syllabes métriques	d ə	s ɛ	t ɛ̯	n ɥ i	f e	n i s	a	t y	v y	l a	s p l ɑ̃	d œ r
Voyelles métriques	ə	ɛ	ə ɛ̯	ɥ i	e	i	a	y	y	a	ɑ̃	œ
Repérage des fins de mots (fdm)	fdm	-	fdm	fdm	-	fdm	fdm	fdm	fdm	fdm	-	fdm
Catégories syntaxiques	Prépo-sition	pronom., dét.	pronom., dét.	nom	nom propre	nom propre	verbe	pronom., dét.	adjectif	pronom., dét.	nom	nom
Marquage accentuel (accent)	-	-	-	accent	-	accent	-	accent	accent	-	-	accent
Nb syllabes	12											

Les productions du métromètre mises sous forme de base de données peuvent être explorées dans plusieurs directions : construction de la figure globale du vers (phonétique, syntaxique, accentuelle) par pièce, genre, auteur… ; recherche de vers répondant à certaines contraintes (vers constitués de très peu de mots, avec peu ou beaucoup de variations phonétiques…) ; recherche de corrélations entre marquages (liens entre le contenu lexical et la forme métrique, entre le genre et la rime…). Nous proposons une nouvelle exploitation de la base de vers enrichie de traits de description pour explorer les questions d'attribution.

2. Les contours de la syllabe comme marque d'auteur

L'application du métromètre à l'ensemble des pièces permet d'identifier des variations dans le traitement métrique. Nous avions déjà observé pour Corneille comme pour Racine une très grande cohérence dans la définition des syllabes métriques. Un mot donné était traité de la première à la dernière pièce de la même manière. Il faut rappeler que Corneille a procédé en 1660 à une révision de ses premières pièces. Dans les trente années qui séparent sa première pièce de 1660, la prosodie a connu des évolutions notables, qui ont été prises en compte dans ses révisions. Celles-ci contribuent à donner le sentiment d'une grande cohérence métrique de bout en bout, quel que soit le genre. Par ailleurs, nous avions noté une seule différence entre Corneille et Racine liée au traitement métrique de *hier* : pour Corneille, il constitue une syllabe, tandis qu'il en représente deux pour Racine. Comment se situe Molière par rapport au traitement de la syllabe métrique ?

Pour ce faire, nous avons sélectionné tous les vers qui d'après le métromètre étaient constitués de 11 ou 13 syllabes. Ces vers constituent forcément des erreurs ou variations de versification puisqu'il n'y a aucun vers de ces longueurs là dans le corpus. Cette sélection permet d'identifier les mots dont l'analyse phonético-métrique se distingue de celle de Corneille et Racine, ces auteurs ayant servi à étalonner le métromètre. Ensuite, sont extraits tous les vers qui contiennent ces mots. Ainsi, peuvent être identifiées d'éventuelles variations dans le traitement métrique, entre auteurs ou chez un même auteur.

Molière contrairement à Corneille, ne semble pas s'être soucié de réviser ses pièces. Ainsi, trouve-t-on dans les deux premières pièces en vers *L'étourdi* (1653) et le *Dépit amoureux* (1656) des cas du type :

> Comme vous voudriez bien, manier ses ducats ;
>
> L'Étourdi ou Les Contre-temps, acte I, scène II.
>
> Et vous devriez mourir d'une telle infamie.
>
> Dépit amoureux, acte V, scène VII.

Où *voudriez* et *devriez* comptent pour deux syllabes. C'est au cours du XVIIème que le décompte des séquences consonne+liquide+I+V se transforme : le *i* acquiert une valeur vocalique et un décompte avec diérèse devient la règle (*de-vri-ez)*. Après ces deux pièces, le nouveau décompte sera adopté par Molière.

Le traitement de *oui* et *hier* nous paraît particulièrement intéressant. Ces deux mots partagent une double particularité : ils peuvent constituer une ou deux syllabes métriques ; entraîner ou non l'élision du *e* muet qui les précède. Ils pourraient favoriser une forme d'élasticité du vers. Corneille et Racine ne jouent pas de cette élasticité potentielle : ni l'un ni l'autre ne varient dans le traitement de ces mots.

Commençons par *oui*. Chez les trois auteurs, *oui* est traité comme un monosyllabe. En cela, ils optent pour une prononciation moderne de *oui*. En effet, en ancien français, le *oui* était dissyllabique. D'après Elwert (1965), le *ou* devant voyelle acquiert une valeur consonantique à partir du milieu du XVIIème siècle. Cependant le *ou* n'est pas traité de la même manière selon les auteurs. Alors que pour Corneille et Racine, un *e* muet s'élide toujours devant *oui* (le *ou* a donc une valeur vocalique), il est tantôt maintenu, tantôt élidé chez Molière.

Ainsi trouve-t-on dans la même pièce

> Notre soeur est folle, oui. / Cela croît tous les jours. (*e* de *folle* élidé devant *oui*, comme chez Corneille et Racine)
>
> Les Femmes savantes, II, IV.
>
> Moi, ma mère ? / Oui, vous. Faites la sotte un peu. (*e* final de *mère* maintenu)
>
> Les Femmes savantes, III, IV.

Sur 38 vers comprenant *oui* après un *e* muet, le *e* muet est élidé 29 fois et maintenu 9 fois. Les cas où le *e* muet est maintenu se situent plutôt dans les comédies les plus burlesques. Le traitement métrique de *oui* est donc instable chez Molière.

Le décompte de *hier* distingue Racine, qui le traite comme un mot dissyllabique de Corneille et Molière qui considèrent le mot comme monosyllabique. Pour Corneille et Racine, le *h* de hier est non aspiré, tout comme chez Molière. Cependant à deux reprises, Molière hésite (ces vers ont été vérifiés dans l'édition d'origine et dans celle du XIXème siècle) :

> Et c'est l'homme qu'hier vous vîtes du balcon. (diérèse sur *hier*)
>
> L'École des femmes, II, V.
>
> Et non comme témoin de ce que hier vous vîtes. (*h* de *hier* aspiré)
>
> Dépit amoureux, II, VI.

Enfin, le mot *biais*, absent chez Corneille, traité avec diérèse chez Racine, voit son traitement varier chez Molière : six cas où *biais* est dissyllabique, deux cas où il est monosyllabique.

> Et vous deviez chercher quelque biais plus doux. (*biais* dissyllabique)
>
> Molière, Le Tartuffe ou L'Imposteur, V, I.
>
> Voyons, voyons un peu par quel biais, de quel air, (*biais* monosyllabique)
>
> Molière, Le Misanthrope, IV, III

A travers l'étude des variations dans la manière de traiter la syllabe métrique, il apparaît que chez Molière, le traitement du vers est plus relâché et moins cohérent que chez Corneille et Racine. Tandis que ces derniers sont cohérents de bout en bout sans exception, Molière hésite parfois et fait varier le traitement de trois mots : *oui*, *hier*, *biais*. Seule une exploration systématique pouvait permettre d'identifier *tous* les lieux où il existe bel et bien une variation de traitement de la syllabe. Il s'agit cependant de phénomènes qui ne concernent qu'un faible, voire très faible nombre de vers. Il est donc nécessaire de déployer une approche plus globale.

3. Marqueurs morpho-syntaxiques et accentuels

En nous limitant aux comédies des trois auteurs, nous avons cherché à voir si la répartition des marquages morphosyntaxiques et accentuels produits par le métromètre permettait de différencier les auteurs. Force est de constater qu'il n'y a pas de délimitation nette entre les comédies selon les auteurs si l'on examine un à un les différents critères. Seule *Les Plaideurs* de Racine se distingue nettement de toutes les autres comédies : beaucoup moins d'hémistiches réguliers, moins de mots-outils en début d'hémistiche. Molière se situe entre Racine et Corneille : les hémistiches "réguliers" avec des accents en position paire (type 010101) ou multiple de trois (001001) y sont plus fréquents que chez Racine mais moins que chez Corneille : on ne peut cependant définir des seuils de ruptures. Certains indicateurs marginaux permettent de distinguer les auteurs. Il en est ainsi de "Et" en début de vers. Plus de 17% des vers de Molière commencent par "Et", contre 11% en moyenne des pièces de Corneille et 7% des *Plaideurs*.

La répartition des différents indicateurs ne paraît pas être un critère sûr, puisque tous les indicateurs ne convergent pas dans le même sens, et qu'on ne peut définir des seuils. Dès que l'on descend au niveau des pièces (voir annexe), on observe que certaines pièces de Corneille se trouvent dans la zone de Molière et inversement, et cela quels que soient les critères observés. Des effets de sous-genre expliquent sans doute ces variations.

Comédies	Nbre vers	H1 001001	H1 010101	H1 autre	H2 001001	H2 010101	H2 autre	"Et" début vers	Mots-outils en p1	Mots-outils en P2
Corneille	13296	3084	4806	5406	3867	5192	4237	1444	9209	5266
		23%	36%	41%	29%	39%	32%	11%	69%	40%
Molière	17058	3741	5675	7642	4867	6190	6001	2576	11338	7082
		22%	33%	45%	29%	36%	35%	15%	66%	42%
Racine	871	139	293	439	203	316	352	42	404	258
		16%	34%	50%	23%	36%	40%	5%	46%	30%
Total	31225	6964	10774	13487	8937	11698	10590	4062	20951	12606
		22%	35%	43%	29%	37%	34%	13%	67%	40%
		chisq =44 ; P<0,0001			chisq=27 ; P<000,1			chisq=210 ; P<0,001	chisq=49 ; P<0,0001	chisq=82 ; P<0,0001

Clef de lecture : Sur les 13296 vers de Corneille, 3084, soit 23%, ont un premier hémistiche de la forme 001001

Tableau 2. Répartition de quelques critères linguistiques pour les comédies selon les auteurs

Labbé (2001) a montré que *le Menteur* et *la suite du Menteur* de Corneille présentaient de fortes parentés lexicales avec les comédies de Molière. Cette proximité lexicale, confirmée par d'autres méthodes, se retrouve au niveau métrique sur quelques rares critères comme la part de mots-outils en première position, plus élevée que dans les autres pièces de Corneille :

il n'y a donc pas forcément de convergence entre les traits liés au vocabulaire et les traits métriques, ces derniers étant assez divergents selon les pièces. Explorons à présent des approches qui traitent globalement la question du vers, en s'appuyant sur la séquence des syllabes métriques.

4. Modèles statistiques de la syllabe

Dans cette section, nous proposons et évaluons un modèle statistique du vers classique, vu sous la forme d'une séquence de douze syllabes, chacune étant représentée par la suite de phonèmes construite par le métromètre. Cette modélisation statistique vise à répondre à un certain nombre de questions : les différences repérées entre les différents auteurs concernant le fonctionnement de certaines positions métriques peuvent-elles être prise en compte simultanément pour fonder un modèle statistique du vers d'un auteur ? De tels modèles peuvent-il avoir des vertus prédictives pour identifier un genre ou un auteur ? Peut-on en déduire de nouveaux indices permettant de caractériser un type d'écriture ?

Dans un premier temps, nous présentons sommairement le type de modèles probabilistes que nous avons utilisé, ainsi que leur utilisation pour différentes tâches (classification d'un vers ou d'un ensemble de vers). Nous présentons ensuite le corpus de vers sur lequel nous avons travaillé, ainsi que les résultats de diverses expérimentations conduites avec ce corpus.

4.1. Modèles de langage

4.1.1. Bases

Les modèles de langage sont des modèles probabilistes permettant de modéliser des *séquences d'événements* prenant des valeurs parmi un inventaire fini. Ces modèles ont été particulièrement utilisés dans le contexte de la reconnaissance automatique de la parole, et il existe à leur sujet une littérature abondante, en particulier (Jelinek, 1997), auxquels nous renvoyons le lecteur. Dans le cadre de cette étude, nous avons utilisé les plus simples de ces modèles, connus sous le nom de modèles n-grammes. Soit $S=s_1...s_n$ une séquence, où les s_i appartiennent tous à un ensemble fini V (le vocabulaire), un tel modèle décompose $Pr(S=s_1...s_n)$ selon:

$$Pr(S=s_1...s_m) = Pr(s_1)Pr(s_2|s_1)Pr(s_3|s_1s_2)\dots Pr(s_n|s_1\dots s_{m-1}) \quad (1)$$

$$\approx Pr(s_1)Pr(s_2|s_1)Pr(s_3|s_1s_2)...Pr(s_m|s_{m-n+1}...s_{m-1}) \quad (2)$$

Le passage de (1) à (2), qui caractérise les modèles n-grammes, correspond à une approximation selon laquelle la probabilité conditionnelle d'émission du symbole s ne dépend que des $n-1$ symboles précédents. Par exemple, dans un modèle d'ordre 2 (bigramme), la probabilité d'un événement dans une séquence dépend uniquement de l'événement précédent. Cette approximation permet de ne faire intervenir dans le modèle qu'un nombre fini (de l'ordre de $|V|^n$) de paramètres de la forme $Pr(s|h)$, où h est une séquence d'au plus $n-1$ symboles. Un modèle bigramme est ainsi paramétré par $|V|$ distributions $Pr(./s)$, chacune de ces distributions ayant $|V|$ paramètres.

Par exemple pour modéliser une suite de 0 et de 1 (le vocabulaire V est constitué de deux élements, 0 et 1, $|V|=2$) dans un modèle bigramme, il suffit de quatre paramètres ($|V|^2=4$) : la probabilité d'avoir un 0 après 0 ($Pr(0|0)$), 1 après 0 ($Pr(1|0)$) ; un 0 après 1($Pr(0|1)$) et 1 après 1 ($Pr(1|1)$). Il suffit en fait d'estimer deux paramètres, les deux autres se déduisant du fait de la contrainte de sommation à 1.

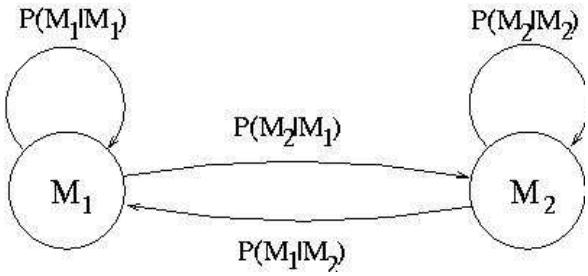
Les cas d'application de ces modèles au traitement automatique des langues, impliquant la modélisation de séquences de parties du discours, de lemmes ou de formes graphiques, imposent de travailler avec des inventaires de grande taille (*V* contenant de quelques centaines à quelques dizaines de milliers de symboles). Même pour de petites valeurs de *n* (*n=2* ou *n=3*), le nombre de paramètres est alors exagérément grand. Le problème d'estimation associé en est rendu d'autant plus difficile que, du fait des répartitions très inégale des occurrences de symboles, la très large majorité des séquences ne sont en fait jamais observées, même dans des corpus de très grande taille. En conséquence, estimer *Pr(s|h)* au maximum de vraisemblance par *Pr(s|h) = n(hs)/n(h)*, avec *n(x)* le nombre d'occurrences de l'événement *x*, conduit à affecter une probabilité nulle à de nombreux événements. De nombreuses techniques de lissage de ces distributions, fondées sur l'interpolation de plusieurs modèles ou des stratégies de repli sont donc, dans la pratique, indispensables (Chen et Goodman, 1996).

4.1.2. Utilisation des modèles stochastiques

Supposons que l'on dispose maintenant de plusieurs sous-corpus $C_1...C_k$, pour lesquels des modèles $M_1...M_k$ ont été construits. Deux tâches peuvent être distinguées : la *classification* qui consiste à attribuer un vers isolé tiré au hasard à un modèle ; la *segmentation* qui revient à distinguer dans un flux de vers, des sous-séquences homogènes.

Pour la classification, il suffit de calculer, pour toute séquence *S* de symboles (une séquence étant un vers), le corpus auquel elle se rattache en évaluant le modèle qui rend cette séquence le plus probable.

Pour la segmentation, qui modélise des séquences de vers, il est également possible de construire des segments homogènes, attribuables à un même modèle M_i. Le modèle de segmentation le plus simple consiste à affecter chaque séquence au meilleur modèle, indépendamment des affectations des vers adjacents ; des modèles plus riches permettent de favoriser les fragments homogènes, en pénalisant les changements de modèles. Par exemple, Le rattachement d'une succession de séquences est alors déterminé par programmation dynamique, en calculant les chemins les plus probables dans des automates stochastiques similaires à celui de la Figure 1, qui capture des dépendances d'ordre 1 entre deux modèles M_1 et M_2. Dans ce modèle, la décision de classement d'un vers dépend du classement du vers précédent. En faisant varier les paramètres de ce modèle, il est possible de favoriser l'homogénéité des segments : en particulier, plus $P(M_1|M_1)$ est grand par rapport à $P(M_2|M_1)$,



plus il sera "difficile" de quitter le modèle M_1.

Figure 1. Modèle d'ordre 1 pour la segmentation des pièces

4.2.Résultats expérimentaux

Nous décrivons ici les résultats de diverses expérimentations conduites en utilisant ces modèles statistiques : nous décrivons d'abord le corpus utilisé, puis nos expériences de classification de vers isolés, puis de classification et de segmentation de groupes de vers.

4.2.1. Description du corpus

Le corpus utilisé contient l'intégralité des alexandrins (du moins ceux reconnus comme tels par le métromètre) des tragédies et des comédies en vers de Corneille, Molière et Racine. Ceci exclut en particulier les vers qui ne sont pas des alexandrins (3% du corpus) ou encore les vers mal analysés (vers en latin ou en patois). Au total, ce corpus comprend 79456 vers, tirés de 51 pièces différentes[iii]. Chaque vers est représenté comme une suite de 12 syllabes. Au total YYY syllabes différentes apparaissent au moins une fois dans ce corpus. Pour les besoins de notre analyse, nous avons extrait intégralement 6 pièces : "Cinna", "Le Menteur", "La suite du Menteur", de Corneille; "Psyché", sur laquelle Corneille et Molière ont travaillé ensemble, et "Les Plaideurs", unique comédie de Racine.

Dans un premier temps, nous décrivons des expériences de classification de vers, qui reposent toutes sur le même protocole, consistant successivement à :

- construire un inventaire des syllabes: dans toutes nos expériences, il s'agit de l'ensemble des syllabes qui ont une fréquence relative supérieure à 10^{-5}. Cet inventaire contient environ 1900 syllabes.
- construire pour chaque sous-corpus un modèle statistique[iv] en utilisant 90% des vers de ce sous-corpus pour l'estimation des paramètres, les 10% restants étant utilisés pour les tests. La répartition entre apprentissage et test est effectuée en tirant au hasard.
- calculer la probabilité de chacun des vers des corpus de test pour chacun des modèles ainsi construit; *attribuer le vers au modèle dans lequel il est le plus probable.*

Nous décrivons ensuite des expériences visant à utiliser nos modèles statistiques pour réaliser la classification de vers ou de groupes de vers selon le procédé décrit à la section précédente.

4.2.2. Classification des vers

La première question à laquelle nous nous sommes intéressés est celle de savoir si de tels modèles probabilistes étaient à même de capturer des éléments de caractérisation des différents auteurs et genre. A cet effet, nous avons effectué des expériences de classification en sous-divisant le corpus (i) par genre et (ii) par genre et auteur; puis en utilisant les modèles appris sur ces sous-corpus comme des outils de classification de vers. Les performances de ces systèmes, mesurées en pourcentage de vers bien classés, sont reproduites respectivement dans les Tables 3et 4

[iii] *Clitandre*, *Don Sanche d'Aragon*, *Andromède*, *La Toison d'or*, *Tite et Bérénice*, *Pulchérie*, toutes de Corneille, ainsi que la très courte *Pastorale comique* de Molière, qui ne relèvent pas exclusivement d'un seul genre, ne figurent pas dans notre corpus.

[iv] Toutes nos expériences ont été menées en utilisant la boîte à outils statistiques SRILM du SRI (Stolcke, 2002). Voir http://www.speech.sri.com/projects/srilm/

	Comédie	Tragédie
Comédie	1519/60.37	657/13.90
Tragédie	997/39.63	4071/86.10
Total	2516/100.00	4728/100.00

Tableau 3. Classification par genre, avec un modèle 3-gramme: 86% des vers tirés des tragédies, soit 4071 vers, sont attribués au modèle de tragédie.

Dans le cas de l'identification des genres, on note principalement que si le système classe majoritairement les vers tirés de tragédies comme tragiques, et ceux tirés de comédies comme comiques, les deux genres sont inégalement bien appris: le modèle de tragédie, appris sur près de deux fois plus de vers, fournit un meilleur catégorisateur que le modèle de comédie.

Dans la deuxième expérience, la sous-catégorisation du corpus conduit à construire 4 modèles, estimés sur des ensembles de taille comparable: un pour les tragédies de Racine, un pour celles de Corneille, un pour les comédies de Corneille, et un pour celles de Molière.

	Molière	Corneille (C)	Corneille (T)	Racine
Molière	944/57.35	150/17.24	236/7.94	140/7.97
Corneille (C)	134/8.14	333/38.28	252/8.48	60/3.41
Corneille (T)	410/24.91	297/34.14	1720/57.89	594/33.81
Racine	158/9.60	90/10.34	763/25.68	963/54.81
Total	1645/100.00	869/100.00	2970/100.00	1756/100.00

Tableau 4. Classification par genre et auteur, avec un modèle 2-gramme: 10.3% des vers tirés des comédies de Corneille, soit 90 vers, sont attribués au modèle de Racine.

Ainsi que ces résultats le montrent sans ambiguïté, les modèles statistiques de type n-gramme capturent des régularités statistiques qui permettent de reconnaître, avec une précision très supérieure au hasard, le genre ou l'auteur de chaque vers de notre corpus. Rappelons que, dans ces expériences, la classification est opérée vers par vers : en utilisant le résultat du classement d'une successions de vers, il devient alors possible d'envisager de rattacher avec très grande précision un groupe de vers à un genre ou à un auteur.

4.2.3. Classification de pièces

Les expériences décrites dans cette section visent à répondre à la seconde question, celle de l'attribution d'un groupe de vers à un genre ou un auteur. Pour tenter d'y répondre, nous avons utilisé les pièces initialement mises de côté: pour chacune, nous avons utilisé les modèles d'auteurs construits précédemment pour classer chacun des vers: dans un premier temps les classements sont effectués de manière indépendante; dans un deuxième temps on pénalise les changements d'auteurs, afin de favoriser les fragments "homogènes", c-a-d attribués à un même auteur, en utilisant un modèle d'ordre 2.

Premier exemple: "Cinna", tragédie Cornélienne par excellence. Pour cette pièce, l'agrégation simple des résultat des classifications individuelles est sans appel: 942 vers, soit plus de la moitié des 1742 vers de la pièce, sont attribués à Corneille [T], Racine se voyant attribuer plus de la moitié (492) des vers restants. Si l'on pénalise les changements d'auteurs, le résultat est encore plus écrasant: 1321 pour Corneille [T], et 322 pour Racine (voir la Figure 2, 0 correspond au premier cas, 1 au second – pénalisation des changements).

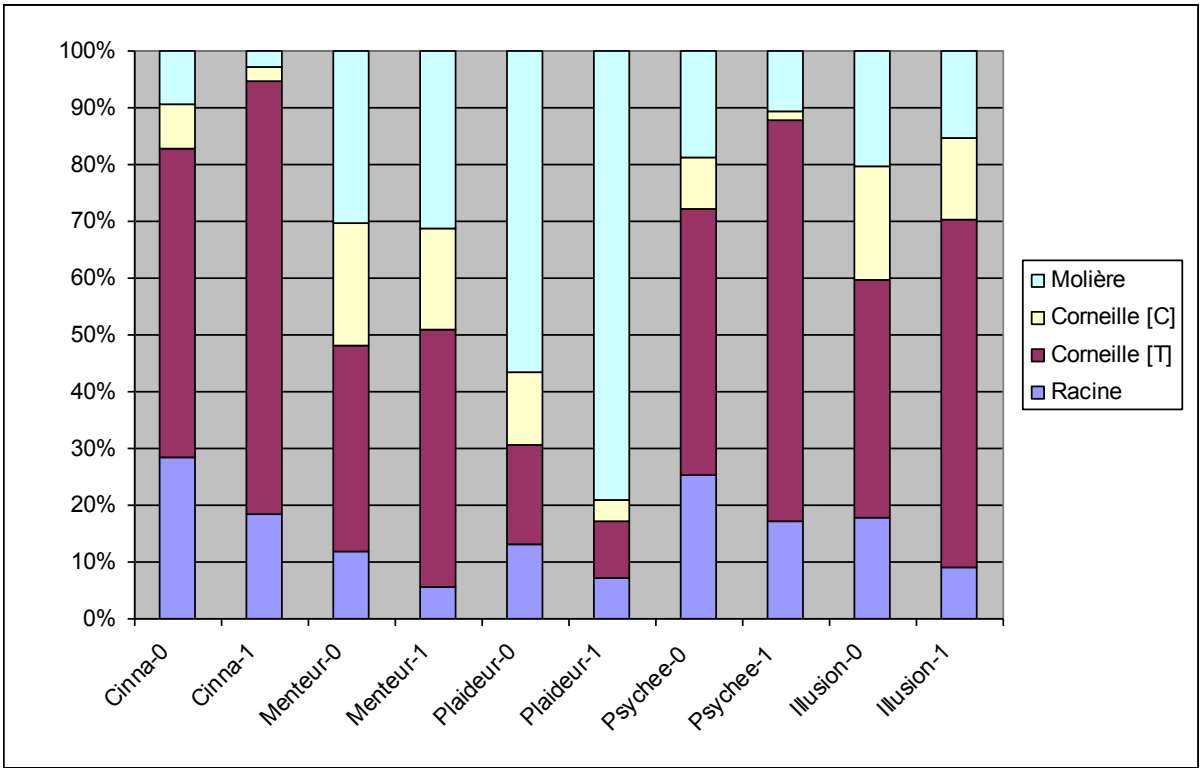


Figure 2. Répartition des vers par modèle pour l'ensemble des pièces

L'utilisation de ce même traitement aux autres pièces de notre corpus permet en particulier de constater (i) la position particulière du Menteur (et de la "Suite du menteur", non représentée ici), qui bien que majoritairement attribué à Corneille, présente un fort pourcentage de vers attribué à Molière; (ii) la proximité des "Plaideurs" avec les pièces de Molière: on retrouve ici la grande différence métrique entre les tragédies de Racine et cette unique comédie; (iii) le partage en genre de "l'Illusion comique" (un tiers Comédie, deux tiers tragédie), attribué sans hésitation à Corneille. En ce qui concerne finalement "Psyché", nous avons pu vérifier que lorsque l'on restreint le modèle de classification à ne choisir qu'entre Corneille et Molière, alors on retrouve une segmentation conforme à ce qu'on sait de l'écriture de la pièce: les vers attribués à Molière sont très massivement localisés dans le début de la pièce[v].

Les résultats obtenus précédemment semblent confirmer l'hypothèse qu'il existe des manières différentes de composer les vers, qui varient suivant les auteurs et les genres (ainsi naturellement à travers les périodes, bien que ce point n'ait pas été abordé ici), et que ces différences peuvent être capturées par des modèles statistiques. Ces résultats ne permettent

[v] "Ainsi, il n'y a que le Prologue, le premier acte, la première scène du second, et la première du troisième, dont les vers soient de lui [Molière]. M. Corneille a employé une quinzaine au reste" Le libraire au lecteur

pas pour autant de conclure définitivement sur la question de l'attribution : rappelons que l'organisation même des corpus d'apprentissage pose comme préalable que ces différences existent ; la modélisation statistique et les expériences ici présentées ne validant ces hypothèses que sous une forme quasi-tautologique: "si ces différences existent, alors elles peuvent être capturées par des modèles statistiques". Pour répondre plus catégoriquement à ces questions, il faudrait être à même de proposer (et de modéliser) une hypothèse "nulle", consistant à supposer que de telles différences n'existent pas, et à comparer ces deux hypothèses. Ce travail reste encore à faire, mais du moins espérons-nous avoir montré qu'il valait la peine d'être entrepris.

Conclusion

Les aspects métriques constituent des indices dont il est nécessaire de tenir compte dans les travaux de stylométrie. La manière de procéder au décompte des mots peut par exemple être un indice pertinent pour distinguer des auteurs. Pour la première fois, les modèles de langage (modèles probabilistes de séquences d'événements) sont appliqués pour modéliser la séquence des douze syllabes dans le vers pour différents auteurs et genres. Construits sur les pièces les plus typiques de genres et d'auteurs, ils s'avèrent fournir des modèles efficaces pour la tâche d'affectation : les résultats trouvés semblent montrer qu'il existe bel et bien des modèles de vers différents selon les genres et les auteurs.

Références

Baayen H., Halteren H. v., Neijt A. & Tweedie F. (2002). An experiment in authorship attribution. JADT 2002 : 6èmes Journées internationales d'Analyse statistique des Données Textuelles, Saint-Malo, 335-346.

Beaudouin V. & Yvon F. (1996). "The Metrometer : a Tool for Analysing French Verse", Literary & Linguistic Computing, vol. 11, n°1, p. 23-32.

Beaudouin V. (2002). Mètre et rythmes du vers classique - Corneille et Racine. Paris, Champion, coll. Lettres numériques.

Bernet C. (1983). Le vocabulaire des tragédies de Racine. Paris-Genève, Slatkine-Champion, 385 p.

Chen S. and Goodman J. (1996). An Empirical study of smoothing techniques for language modeling In the Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pages 310-318, Santa Cruz, NM.

Constant P. (1991). Analyse syntaxique par couche, Doctorat ENST, Paris, ENST.

Elwert W. T. (1965). Traité de versification française. Des origines à nos jours. Paris, Klincksieck, 210 p.

Holmes D. I. (1998). "The Evolution of Stylometry in Humanities Scholarship", Literary and Linguisitic Computing, Vol. 13, n°3, p. 111-117.

Jelinek F. (1997). Statistical Methods for speech recognition. The MIT Press, Cambridge, CA.

Labbé C. & Labbé D. (2001). "Inter-Textual Distance and Authorship Attribution Corneille and Molière", Journal of Quantitative Linguistics, Vol. 8, n°3, p. 213-231.

Muller C. (1967, 1992). Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille. Paris, Larousse, 1967, réimpression aux éditions Slatkine, 1979, 1992 382.

Stolcke, A. (2002). SRILM -- An Extensible Language Modeling Toolkit. In the Proceedings of the International Conference on Speech and Language Processing, pages 901-904, Denver, CO.

Yvon F. (1996). Prononcer par analogie : motivation, formalisation et évaluation, Thèse de l'ENST, 278 p.

Annexe

		Nb d'alex-andrins	Mots-outils en P1	Mots-outils en P7	"Et" début vers	h1 001001	h1 010101	h1 autre	h2 001001	h2 010101	h2 autre
R	Les Plaideurs	645	63%	64%	7%	17%	35%	48%	25%	37%	38%
C	Mélite	1735	67%	73%	9%	24%	36%	40%	32%	37%	31%
C	La Galerie du palais	1708	67%	72%	9%	20%	39%	42%	30%	39%	32%
C	Le Menteur	1676	72%	71%	10%	22%	34%	44%	26%	40%	34%
M	L'Étourdi ou Les Contre-temps	1875	68%	73%	10%	22%	34%	44%	28%	38%	34%
C	La veuve	1827	67%	74%	11%	23%	37%	40%	29%	39%	32%
C	La Suivante	1602	69%	74%	11%	25%	36%	39%	30%	40%	31%
C	L'Illusion comique	1600	69%	75%	12%	25%	35%	41%	29%	40%	31%
C	La Suite du menteur	1714	73%	75%	12%	23%	36%	41%	27%	40%	33%
C	La Place royale	1434	71%	75%	13%	25%	37%	38%	31%	37%	32%
M	Sganarelle ou Le Cocu imaginaire	599	70%	73%	13%	22%	36%	42%	34%	35%	32%
M	Dépit amoureux	1583	71%	72%	14%	22%	32%	46%	30%	36%	33%
M	Les Fâcheux	760	69%	70%	15%	24%	33%	43%	28%	36%	35%
M	L'École des maris	986	73%	72%	16%	24%	31%	45%	31%	37%	32%
M	L'École des femmes	1584	72%	71%	17%	22%	32%	46%	28%	36%	36%
M	Dom Garcie de Navarre ou Le Prince jaloux	1813	70%	73%	17%	25%	35%	41%	33%	36%	31%
M	Amphitryon	954	72%	72%	18%	22%	32%	46%	30%	37%	33%
M	Les Femmes savantes	1611	75%	73%	18%	22%	35%	43%	29%	37%	34%
M	Le Misanthrope	1676	73%	73%	18%	21%	34%	45%	27%	37%	36%
M	Mélicerte	556	72%	74%	19%	21%	33%	46%	32%	36%	32%
M	Le Tartuffe ou L'Imposteur	1808	75%	75%	20%	21%	36%	43%	27%	38%	35%

Tableau 5. Répartition de quelques critères linguistiques pour les comédies